ABSTRACT OF THE DISCLOSURE

An apparatus and method for determining if a query document matches one or more of a plurality of documents in a database. In a coarse matching stage, a compressed file or other query document is scanned to produce a bit profile. Global statistics such as line spacing and text height are calculated from the bit profile and used to narrow the field of documents to be searched in an image database. The bit profile is cross-correlated with bit profiles of documents in the search space to identify candidates for a detailed matching stage. If multiple candidates are generated in the coarse matching stage, a set of endpoint features is extracted from the query document for detailed matching in the detailed matching stage. Endpoint features contain sufficient information for various levels of processing, including page skew and orientation estimation. In addition, endpoint features are stable, symmetric and easily computable from commonly used compressed files including, but not limited to, CCITT Group 4 compressed files. Endpoint features extracted in the detailed matching stage are used to correctly identify a matching document in a high percentage of cases.